Exercise 7

Task description

Train the policy for a pybullet locomotor problem with the PPO algorithm (do not forget to restore the original pybullet files with the reward functions suitable for reinforcement learning, as described in Section 3.3). Compare the result with those obtained with evolutionary strategies.

Solution

For the exercise the hopper model was used for training and analyzing purposes. The hopper is a two-dimensional one-legged robot that after the jump should balance.

On figure 1, there are the results of training of a hopper model on the custom reward. As seed for training it is used Seed = 3.

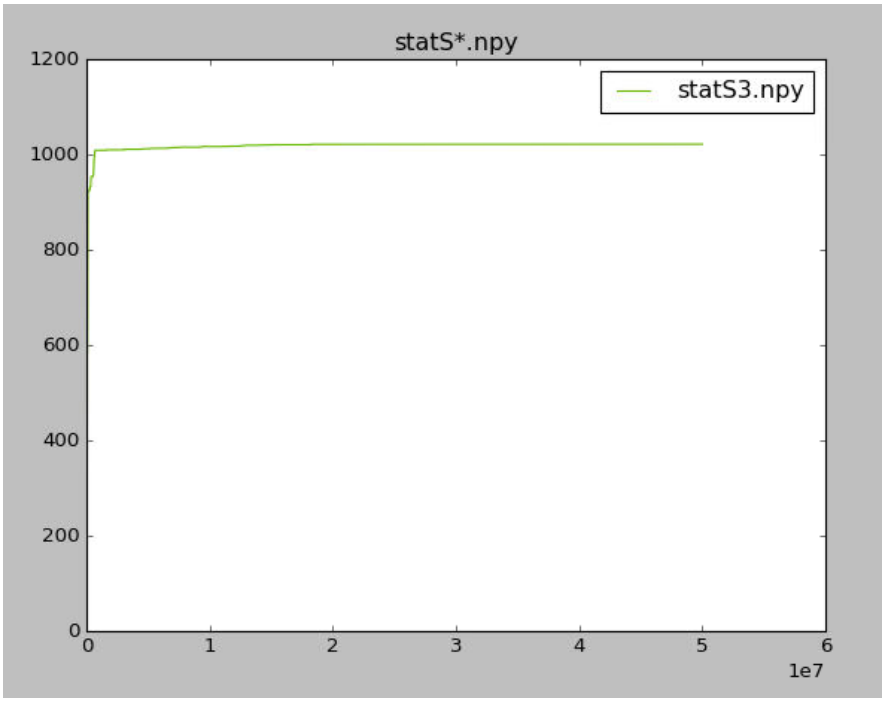

Figure 1: The variations of the performance across generations

The average and the standard deviation of performance among multiple runs:

- Custom reward function:

  Average Generalization: 1021.52 +-0.00

- Original reward function:

  Average Generalization: 1036.14 +-0.00

Then the hopper is trained with baselines library with PPO2 algorithm. The result is presented on the figure 2.

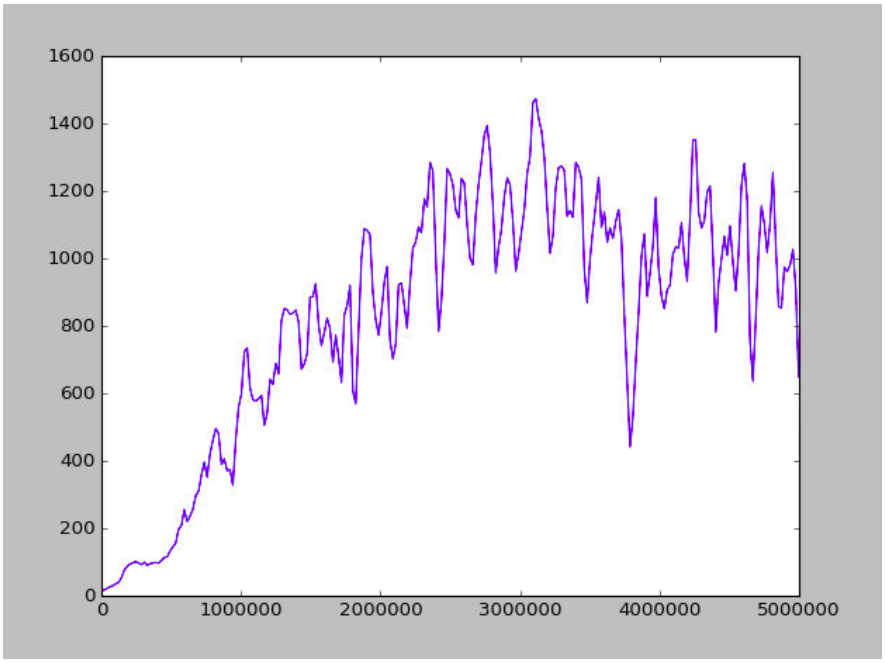

Figure 2: The variations of the performance across generations using PPO2 algorithm

Comparing the two figures, it can be noticed that training with evolutionary strategies gives more stable reward across multiple runs that increase on the certain amount and reaches more than 1000 reward value. The training with PPO2 algorithm gives not stable progress in reward function, which unstableness starts approximately from 600 and varies around 1000 value. The maximum reward value almost reaches 1500. Maybe to train the model with different seed and network will give more stable reward progression.